微博用户兴趣主题抽取方法

杨仁凤[1,2]，陈端兵[1,2]，谢文波[1,2]

(1. 电子科技大学计算机科学与工程学院 成都 611731；2. 电子科技大学大数据研究中心 成都 611731)

【摘要】根据社交媒体短文本特征改进了词袋模型，利用特征之间的语义关系提出了语义表示模型，采用句子中特征先后顺序构建了次序图模型，在此基础上引入时间因素，提出了基于Single-Pass算法的用户兴趣主题模型用于抽取微博用户关注的话题。实验结果表明，该方法的FM、AA和F指标相比FSC-LDA方法分别提高了200.40%、46.50%、80.05%。

关 键 词 兴趣抽取；微博；Single-Pass；文本聚类；主题模型

中图分类号 TP181 **文献标志码** A doi:10.3969/j.issn.100 1-0548.2018.04.025

A Method of Micro-Blog Users′ Interests Topic Extraction

YANG Ren-feng[1, 2], CHEN Duan-bing[1, 2], and XIE Wen-bo[1, 2]

(1. School of Computer Science and Engineering, University of Electronic Science and Technology of China Chengdu 611731;
2. Big Data Research Center, University of Electronic Science and Technology of China Chengdu 611731)

Abstract The bag of word model is first improved according to the social media short text feature. The semantic representation model is then proposed by using semantic relations between features. The sequence diagram model can be constructed by using the sequence of features in the sentence. On the base of these, together with time factor, we propose a user interest topic mode based on Single-Pass to extract the topic of user's attention. The experimental results show that the FM, AA and F of our method are increased by 200.40%, 46.50% and 80.05%, respectively, compared with the latest method FSC-LDA.

Key words interests extraction; micro-blog; single-pass; text clustering; topic model

随着大数据时代[1]的来临，微博作为一种融合短信息、社交网络和传播媒体的工具和平台，吸引了成千上万的用户[2]。微博具有流行性、及时性及交互性等特点，在某种程度上反映了用户的兴趣。由此衍生出新的商业应用及社会现象，如广告推送[3,4]、网络舆情[5-9]等。

为了充分挖掘社交媒体的价值，研究者们提出了大量用户兴趣挖掘方法。其中，利用用户发表的微博内容构建主题模型挖掘用户兴趣是一种常见方法。文献[10]利用文本信息中的主题标签、用户行为和时间来挖掘用户的短期和长期兴趣，从而进行个性化主题推荐。文献[11]从时间、地点、情感、回帖和关系等方面综合分析，揭示动态主题。文献[12-13]利用LDA模型抽取用户兴趣特征，从而构建用户的兴趣主题。文献[14]利用带有标签的LDA模型在Twitter上推断用户兴趣的主题。文献[15]提出在微博上利用改进的LDA模型(FSC-LDA)来监测热点话题。文献[16]探究在社交媒体中利用高斯混合模型对用户兴趣进行预测。

但传统的基于主题的用户兴趣挖掘方法计算复杂度高，在大规模数据上时效性差。因此本文提出一种以Single-Pass[17]聚类算法为核心的用户兴趣主题抽取方法来应对海量流式数据。Single-Pass算法在话题发现与追踪方面使用范围广、适用性强，表现出卓越的聚类性能，其流式数据载入的特性非常适合大规模实时更新的社交媒体数据的分析。但传统方法中将Single-Pass聚类算法与TF-IDF特征提取算法结合的做法仅能处理长文本，对于特征稀疏的社交媒体短文本数据适用性差。

本文在向量空间模型的基础上，结合特征之间的语义相关性和顺序关系，引入时间因素对用户兴趣的影响，提出了基于Single-Pass算法的用户兴趣主题模型，从而实时准确地对用户关注的话题信息进行抽取。其中，特征提取包含3部分：1) 根据短文本特征提出了改进的词袋模型(improved bag of word, IBOW)；2) 利用特征之间的语义关系提出了

收稿日期：2017－02－01；修回日期：2017－08－15
基金项目：国家自然科学基金(61433014，61673085)；中央高校基本科研业务费专项资金(ZYGX2014Z002)
作者简介：杨仁凤(1989－)，男，主要从事数据挖掘、大数据方面的研究.

语义表示模型(Word2Vec, W2V)；3) 采用句子的顺序关系提出了次序图模型(sequence diagram model, SDM)。在此基础上综合考虑IBOW、W2V和SDM来提高Single-Pass的聚类效果。实验结果表明本文方法的FM、AA和F指标相比方法FSC-LDA分别提高了200.40%、46.50%、80.05%。综上，本文提出的微博用户兴趣主题抽取方法能有效地挖掘用户的兴趣。

1 基于Single-Pass算法的用户兴趣主题模型

兴趣主题抽取的目的是提取微博用户关注的话题。但微博数据量大、话题数量不确定、传统的主题模型方法难以实现高效实时的兴趣挖掘。因此，本文采用流式数据的经典聚类算法Single-Pass抽取用户关注的话题。此外，为了更好地应对社交媒体中的短文本数据，本文对文本表示与特征提取、文本相似度计算、质心更新等方面进行了改进，从而提高聚类效果。基于Single-Pass算法的用户兴趣主题模型主要包括数据预处理、文本表示与特征提取、文本相似性评估以及主题构建这4个步骤。

1.1 数据预处理

微博数据格式繁杂多样，具有很强的噪声干扰，在进行用户兴趣挖掘之前需要对微博数据进行加工处理。预处理过程包括：过滤与主题语义无关的信息，分词并对部分特征词进行过滤。

原始微博数据中很多与主题语义无关的信息需要过滤，比如超链接URL。另外，“^_”、“!!”、“www.”等大量的符号字符也需要清理过滤，否则对后面的特征词提取与文本表示将造成严重的噪声干扰。

在微博文本内容进行分词过程中，需要建立中英文的停用词词典文件，对停用词进行过滤，对特征进行降维处理。因此对微博内容进行停用词的处理，能够降低特征信息的维度及停用词带来的噪声影响。

1.2 文本表示与特征提取

文本表示模型是指从文本中抽取相应特征来描述文本内容的方式，特征可以是字、词、短语等形成的向量、树等结构。文本表示主要有布尔模型[18]、概率模型[19]、语言模型[20]、向量空间模型[21]等。定义不同的特征权重计算方法与相似度计算方法，是文本表示模型的核心。

向量空间模型[21]是目前文本表示使用最多的模型之一。向量空间模型基于词袋思想，将文本表示为特征集合从而实现用向量表示文本。向量空间模型中存在两个理论上的假设，一个是特征之间的相对独立性，另一个是特征之间的无序性，这两个假设是对非结构化文本数据的一种简化。本文在向量空间模型的基础上，考虑特征之间的语义相关性和特征之间的顺序关系，改进文本表示。

微博内容属于短文本，经过数据预处理后得到的特征词较少。因此，不需要进行降维处理，可直接把微博内容预处理后得到的所有词作为该篇微博的特征词。

特征词权重计算一般采用经典的TF-IDF算法[21]。TF-IDF算法综合考虑了词在文本中的代表性以及区分度。词在文档中出现的次数越多，说明该词对文本的中心思想表达所做的贡献越大，越具有代表性。而逆文本频率越大，说明该词在很多篇文档中都出现过，越没有区分度。具体采用下式计算：

$$W(w,d)=\frac{\mathrm{tf}(w,d)\times\log\left(\dfrac{N}{\mathrm{df}(w,D)}+0.01\right)}{\sqrt{\sum_{w\in d}\left\{\mathrm{tf}(w,d)\log\left(\dfrac{N}{\mathrm{df}(w,D)}+0.01\right)\right\}^2}} \tag{1}$$

式中，$\mathrm{tf}(w,d)$ 表示词w在文档d中出现的次数；$\mathrm{df}(w,D)$ 表示词w在文档集D中出现的文档数；N表示文档总数，数值0.01表示对log0的处理。

但在实际应用中，词在文本中表现出的不同性质对文本主题的表达具有不同的贡献，因此本文在传统TF-IDF算法基础上，考虑词性、文本归一化处理以及微博短文本的特征，对传统的的TF-IDF算法进行了改进，得到改进的词袋模型(IBOW)。短文本长度短，特征词之间的频数差异小，很难看出哪些特征词更重要。传统TF-IDF算法针对特征词出现在大部分长文本中，表示该特征越没有区分度，会做惩罚处理。而短文本中是基于每个特征词出现在大部分文档中，说明这个特征词越重要，越能体现用户的主题特征。因此，在IBOW中增加了一个鼓励权重$\dfrac{\mathrm{df}(w,D)}{N+0.01}$，其计算公式如下：

$$W'(w,d)=\frac{\lambda\times\mathrm{tf}(w,d)\log\left(\dfrac{N}{\mathrm{df}(w,D)}+0.01\right)+\dfrac{\mathrm{df}(w,D)}{N+0.01}}{\sqrt{\sum_{w\in d}\left\{\lambda\times\mathrm{tf}(w,d)\log\left(\dfrac{N}{\mathrm{df}(w,D)}+0.01\right)+\dfrac{\mathrm{df}(w,D)}{N+0.01}\right\}^2}} \tag{2}$$

式中，特征词w在文档d中词性加权系数λ定义为：

$$\lambda=\begin{cases}1.5 & \text{词}w\text{为命名实体}\\ 1.0 & \text{词}w\text{为其他名词或者动词}\\ 0.5 & \text{其他}\end{cases} \tag{3}$$

由此，在IBOW中利用式(2)可获得每条微博内容的特征词 $w_1,w_2,\cdots,w_m$ 对应的权重 $\mathrm{wt}_1,\mathrm{wt}_2,\cdots,\mathrm{wt}_m$，每条微博内容文本 d_i 按照向量空间模型可表示为：

$$\boldsymbol{V}_{1,d_i}=(\mathrm{wt}_1,\mathrm{wt}_2,\cdots,\mathrm{wt}_m) \tag{4}$$

在向量空间模型的基础上，本文还考虑了特征之间的语义相关性。利用Google的Word2Vec[22-23]模型(W2V)建立词间的语义关系。该模型采用三层神经网络进行构建，将单词映射到向量空间上进行表示。由此，利用W2V模型可将每条微博内容特征词映射成向量 $\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_m$，得到每条微博内容文本 d_i 的语义表示模型：

$$\boldsymbol{V}_{2,d_i}=(\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_m) \tag{5}$$

式中，$\boldsymbol{v}_j=(P_{j,1},P_{j,2},\cdots,P_{j,n})$。

为了更准确地描述语义，本文还利用句子出现的先后关系构建了次序图模型(SDM)。该模型对用户所有微博内容中的每条信息进行断句处理，形成一个以句子为分析单元的集合。对于集合中的每个句子，首先提取句子特征(词或者字)，并将句子转化为特征序列。然后，对句子内部出现的特征，依照先后顺序两两之间连接一条边。边的权重为特征在同一个句子内部共现的频率。通过以上处理，可以构成一个以特征为节点，特征间共现频率为权重的次序图模型。由此，SDM模型可用一个三元组 $G=(V_3,E,W_3)$ 表示，其各项含义如下：

$V_3(G)$：次序图模型的非空有限结点集合。结点集合 $V_3(G)=\{w_1,w_2,\cdots,w_m\}$，其中每个结点对应于一个特征词项。

$E(G)$：次序图模型的有向边集合，表示结点的有序对。

$W_3(G)$：次序图模型边的权重集合 $W_3(G)=\{\mathrm{gt}_1,\mathrm{gt}_2,\cdots,\mathrm{gt}_t\}$，其中 gt_i 为有向边的权重，与特征间共现频率有关。

1.3　文本相似性评估

传统文本表示模型大多基于向量空间模型，相似度计算一般采用欧式距离、余弦相似度以及杰卡德(Jaccard)相似度等算法。本文在向量空间模型上，结合语义表示模型和次序图模型提出了一种混合模型。在此基础上，定义了一种混合相似度计算方法。文档 d_i 和文档 d_j 的相似度采用下式计算：

$$\begin{aligned}\mathrm{Sim}(d_i,d_j)=&\alpha\mathrm{Sim}_{V_1}(d_i,d_j)+\\ &\beta\mathrm{Sim}_{V_2}(d_i,d_j)+(1-\alpha-\beta)\mathrm{Sim}_{V_3}(d_i,d_j)\end{aligned} \tag{6}$$

式中，α、β 为权重调节参数 $0\leqslant\alpha+\beta\leqslant1$，用于调节3个模型在混合相似度计算中的重要程度。$\mathrm{Sim}_{V_1}(d_i,d_j)$ 表示两个文档 d_i 和 d_j 基于改进词袋模型(IBOW)上的相似度；$\mathrm{Sim}_{V_2}(d_i,d_j)$ 为基于语义表示模型(W2V)的相似度；$\mathrm{Sim}_{V_3}(d_i,d_j)$ 为基于次序图模型(SDM)的相似度。

$\mathrm{Sim}_{V_1}(d_i,d_j)$ 采用余弦相似性公式计算：

$$\mathrm{Sim}_{V_1}(d_i,d_j)=\frac{\sum_{k=1}^{m}|\mathrm{wt}_{k,d_i}\times\mathrm{wt}_{k,d_j}|}{\sqrt{\sum_{k=1}^{m}\mathrm{wt}^2{}_{k,d_i}}\times\sqrt{\sum_{k=1}^{m}\mathrm{wt}^2{}_{k,d_j}}} \tag{7}$$

式中，wt_{k,d_i} 表示特征词 w_k 在文档 d_i 中的权重；wt_{k,d_j} 表示特征词 w_k 在文档 d_j 中的权重。

本文将每个文档中所有的特征向量的算术平均作为该文档的中心向量。因此得到文档 d_i 和 d_j 的中心向量：

$$\begin{cases}\boldsymbol{V}'_{2,d_i}=(\mathrm{VC}_{1,d_i},\mathrm{VC}_{2,d_i},\cdots,\mathrm{VC}_{n,d_i})=\dfrac{\sum_{k=1}^{m}(P_{k,1},P_{k,2},\cdots,P_{k,n})}{m}\\ \boldsymbol{V}'_{2,d_j}=(\mathrm{VC}_{1,d_j},\mathrm{VC}_{2,d_j},\cdots,\mathrm{VC}_{n,d_j})=\dfrac{\sum_{k=1}^{m}(P'_{k,1},P'_{k,2},\cdots,P'_{k,n})}{m}\end{cases} \tag{8}$$

根据文档 d_i 和 d_j 的中心向量，采用余弦公式计算两个文档 d_i 和 d_j 的相似度：

$$\mathrm{Sim}_{V_2}(d_i,d_j)=\frac{|\sum_{k=1}^{n}\mathrm{VC}_{k,d_i}\times\mathrm{VC}_{k,d_j}|}{\sqrt{\sum_{k=1}^{n}\mathrm{VC}^2{}_{k,d_i}}\times\sqrt{\sum_{k=1}^{n}\mathrm{VC}^2{}_{k,d_j}}} \tag{9}$$

杰卡德(Jaccard)相似度 $\mathrm{Sim}_{V_3}(d_i,d_j)$ 主要用于衡量个体间共同具有的特征是否一致。假设由文档 d_i 和 d_j 建立两个次序图结构 $G_i=(V_{3i},E_i,W_{3i})$ 和 $G_j=(V_{3j},E_j,W_{3j})$，则文档 d_i 和 d_j 的相似性采用下式计算：

$$\mathrm{Sim}_{V_3}(d_i,d_j)=\begin{cases}0 & W_{3i}\cap W_{3j}=\varnothing\\ 1.0 & k|W_{3i}\cap W_{3j}|\geqslant|W_{3i}\cup W_{3j}|\\ \dfrac{k|W_{3i}\cap W_{3j}|}{|W_{3i}\cup W_{3j}|} & \text{otherwise}\end{cases} \tag{10}$$

式中，k是可调节参数，在实际运用中取值为3获取最佳效果。可调参数k实际上弱化了双方的差异，同时放大了双方的共同特征，这在一定程度上改善了微博数据稀疏的问题。

另外，如果两条微博之间的时间跨度越大，属于同一个话题的可能性越低，因此，本文引入时间衰减指数，对式(6)进行修正得到：

$$\mathrm{Sim}'(d_i,d_j)=\mathrm{Sim}(d_i,d_j)\mathrm{e}^{-\frac{|\mathrm{time}_i-\mathrm{time}_j|}{365}} \tag{11}$$

式中，$-\frac{1}{365}$ 为衰减系数，$\mathrm{e}^{-\frac{|\mathrm{time}_i-\mathrm{time}_j|}{365}}$ 表示随时间衰减程度；time_i 和 time_j 分别为文档 d_i 和 d_j 发布的时间。两条微博的相似度按照时间间隔呈指数衰减，符合实际规律：当时间间隔不大时，相似度衰减缓慢；当时间间隔逐渐增加时，相似度会加速衰减，最后接近0。

1.4 主题构建

本文选用Single-Pass聚类算法为核心构建微博用户主题模型，以便能更好地处理互联网中的大规模、非结构化、按时间顺序组织的微博数据。

Single-Pass算法[17]是单通道聚类算法，是流式数据的经典聚类算法。针对依次到达的流式数据，该算法按照数据到达的次序每次只处理一条数据，计算数据与当前所有类的相似程度，将该数据划分到最大相似度的类中或者以该数据为基础创建一个新的类，从而实现流式数据的增量和动态聚类。

通常情况下，Single-Pass聚类算法与TF-IDF特征提取算法结合可以很好地处理长文本聚类，但这种策略不适合特征稀疏的微博短文本数据。因此，本文采用改进的词袋模型(IBOW)、语义表示模型(W2V)与次序图模型(SDM)的组合模型进行文本特征提取与表示，利用式(6)和式(11)进行文本相似度计算。此外，传统方法在聚类过程中需要迭代更新中心向量。但微博短文本数据的特征稀疏，每次迭代更新会造成中心向量很大的偏移，导致聚类效果不佳。而本文使用的Single-Pass聚类只需设置类簇的中心向量(包括词袋中心向量、语义中心向量和次序中心向量3个维度)，以增量的方式进行聚类。其高效性特别适合处理互联网环境中大规模数据。并且，通过更改聚类融合阈值 θ 还可以控制获取不同粒度的话题。

用户 U_i 在时间段 T_j 发布和转发的 m 条微博信息：$\mathrm{Weibo}_{\mathrm{Vec}}=\{\mathrm{weibo}_1,\mathrm{weibo}_2,\cdots,\mathrm{weibo}_m\}$。通过兴趣主题抽取得到该用户关注的 s 个话题：$\mathrm{Topic}_{\mathrm{Vec}}=\{\mathrm{topic}_1,\mathrm{topic}_2,\cdots,\mathrm{topic}_s\}$。则每个话题 topic_i 的兴趣度被定义为：

$$\mathrm{IS}_{\mathrm{topic}_i}=\frac{\sum\limits_{t_i\in T_j}\frac{\mathrm{count}(\mathrm{topic}_i)}{m}\times\mathrm{e}^{-\mu|t-t_i|}}{\sum\limits_{t_i\in T_j}\mathrm{e}^{-\mu|t-t_i|}} \tag{12}$$

式中，$\mathrm{IS}_{\mathrm{topic}_i}$ 表示关于话题 topic_i 的兴趣度值；$\mathrm{count}(\mathrm{topic}_i)$ 代表用户 U_i 在 t_i 时刻发表关于话题 topic_i 的微博条数；μ 表示衰减因子；t表示当前计算时间；$\frac{\mathrm{count}(\mathrm{topic}_i)}{m}$ 表示用户在 t_i 时刻发表的该类话题的微博数目占发布总信息比例；$\mathrm{e}^{-\mu|t-t_i|}$ 表示对于话题 topic_i 兴趣度的衰减程度。

用户的兴趣主题可进一步表示为：

$$\mathrm{Topic}'_{\mathrm{Vec}}=\left\{\left\langle\mathrm{topic}_1,\mathrm{IS}_{\mathrm{topic}_1}\right\rangle,\left\langle\mathrm{topic}_2,\mathrm{IS}_{\mathrm{topic}_2}\right\rangle,\cdots,\left\langle\mathrm{topic}_s,\mathrm{IS}_{\mathrm{topic}_s}\right\rangle\right\} \tag{13}$$

因此，用户 U_i 在时间段 T_j 的兴趣主题向量可以表示为：

$$\boldsymbol{U}_{i,\mathrm{topic}}=(\mathrm{Weibo}_{\mathrm{Vec}},\mathrm{Topic}'_{\mathrm{Vec}},T_j) \tag{14}$$

2 实验结果与讨论

2.1 实验数据

本文实验数据来源于数据堂提供的微博数据[24]，其中包含63 641个新浪微博用户信息，84 168条在2014-05-03−2014-05-11期间12个人工标注主题的微博信息语料库。简单统计信息如表1所示。

表1 微博数据信息

话题	微博数量/条	属性数/个
魅族	3 263	121 708
小米	11 569	498 859
火箭队	6 364	237 101
林书豪	1 514	62 291
恒大	8 080	319 148
韩剧	7 515	314 090
雾霾	5 955	260 380
房价	8 935	447 107
同桌的你	10 886	479 905
公务员	7 572	369 795
贪官	6 835	320 782
转基因	5 625	272 655

2.2 实验评估指标

文本聚类按照聚类评价指标可以分为基于人工判定的指标和基于目标函数的指标[25]。由于实验数据集是带主题标签的，所以本文使用Fowlkes and Mallows指标(FM)[26]、平均准确率(AA)[27-28]以及类F

值[25,29]这些人工判定的指标对模型进行评估。设真实主题标签集 $P=\{P_1,P_2,\cdots,P_s\}$，算法聚类结果集 $C=\{C_1,C_2,\cdots,C_m\}$，其中 P_i 和 C_i 表示一个类簇，s 不一定等于m。任意两个文本 (d_i,d_j) 的实际类别和预测类别为下列4种情况之一：(实际同类，预测同类)、(实际不同类，预测同类)、(实际同类，预测不同类)、(实际不同类，预测不同类)，语料库中各种情况对应的数量分别为a、b、c、d。

由此，FM指标为：

$$\mathrm{FM}=\sqrt{\frac{a}{a+b}}\sqrt{\frac{a}{a+c}} \tag{15}$$

平均准确率AA定义为积极准确率(PA)[27-28]与消极准确率(NA)[27-28]的算术平均值：

$$\mathrm{PA}=\frac{a}{a+c} \tag{16}$$

$$\mathrm{NA}=\frac{d}{b+d} \tag{17}$$

$$\mathrm{AA}=\frac{\mathrm{PA}+\mathrm{NA}}{2} \tag{18}$$

对任何带标签的主题 P_j 和聚类簇 C_i：

$$\mathrm{Precision}(P_j,C_i)=\frac{|P_j\bigcap C_i|}{|C_i|}=\frac{a}{a+b} \tag{19}$$

$$\mathrm{Recall}(P_j,C_i)=\frac{|P_j\bigcap C_i|}{|P_j|}=\frac{a}{a+c} \tag{20}$$

$$F(P_j,C_i)=\frac{2\times\mathrm{Precision}(P_j,C_i)\times\mathrm{Recall}(P_j,C_i)}{\mathrm{Precision}(P_j,C_i)+\mathrm{Recall}(P_j,C_i)} \tag{21}$$

对于每个带标签的主题 P_j：

$$F(P_j)=\max_{i=1,2,\cdots,m}\left\{F(P_j,C_i)\right\} \tag{22}$$

则最终的F值为

$$F=\frac{\sum_{j=1}^{s}|\mathrm{P}_j|\times F(\mathrm{P}_j)}{\sum_{j=1}^{s}|\mathrm{P}_j|}。 \tag{23}$$

式(19)～(23)是计算基于人工标注类F值的系列公式。式(23)表明可以通过评价全局所有的带主题标签类来评价整个聚类结果。

AA指标综合考虑了积极准确率和消极准确率，在实际应用中可以起到一定的作用；而FM和F值指标对聚类结果优劣的整体区分能力比较强[25]。

2.3　实验结果

为了验证本文提出的改进的词袋模型(IBOW)、语义表示模型(W2V)以及次序图模型(SDM)的聚类效果，本文构建了几个相关组合模型进行有效性分析，具体结果如表2所示。

1) 几种相关组合模型的参数配置

在第1组模型中，仅仅利用IBOW对兴趣主题进行抽取，体现在式(7)中 $\alpha=1$，$\beta=0$。在第2组模型中，仅仅利用W2V对兴趣主题进行抽取，体现在式(7)中 $\alpha=0$，$\beta=1$。在第3组模型中，仅仅利用SDM对兴趣主题进行抽取，体现在式(7)中 $\alpha=0$，$\beta=0$。3组模型中的各项聚类的评价指标FM、AA、F的结果如表3所示。

表2　几种相关组合模型

组号	模型	模型说明
1	IBOW	仅仅考虑改进的词袋模型对微博进行文本表示和相似度计算
2	W2V	仅仅考虑特征之间的语义对微博进行文本表示和相似度计算
3	SDM	仅仅考虑特征之间的顺序对微博进行文本表示和相似度计算
4	IBOW+W2V	在改进的词袋模型的基础上，考虑特征之间的语义相关性对微博进行文本表示和相似度计算
5	IBOW+SDM	在改进的词袋模型的基础上，考虑特征之间的顺序关系对微博进行文本表示和相似度计算。
6	W2V+SDM	仅仅考虑特征之间的语义相关性和顺序关系对微博进行文本表示和相似度计算。
7	IBOW+W2V+SDM	在改进的词袋模型的基础上，综合考虑特征之间的语义相关性和顺序关系对微博进行文本表示和相似度计算。

表3　IBOW、W2V、SDM的聚类效果对比

模型	FM	AA	F
IBOW	0.180	0.517	0.287
W2V	0.710	0.741	0.674
SDM	0.136	0.509	0.202

在第4组模型中，结合IBOW和SDM对兴趣主题进行抽取，体现在式(7)中需要满足 $\beta=0$。通过一组实验来观察参数 $\alpha\in[0.1,1)$ 对各项指标值 $\phi(\alpha)$ 的影响，实验结果如图1a所示。从图1a中可以看出，随着 α 的增加，指标AA波动不大，因此 α 的变化对聚类指标AA影响不大。当 $\alpha=0.9$ 时，聚类指标FM、AA、F都达到最优值。因此IBOW与SDM组合的最优参数设置为 $\alpha=0.9,\beta=0$。

在第5组模型中，结合IBOW和W2V对兴趣主题进行抽取，体现在式(7)中需要满足 $\alpha+\beta=1(\alpha,\beta\neq0,1)$。通过一组实验来观察参数 $\alpha\in[0.1,1)$ 对各项指标值 $\phi(\alpha)$ 的影响，实验结果如图1b所示。从图1b可以看出，当 $\alpha=0.1$ 和 $\alpha=0.7$ 时，聚类的FM、AA、F指标具有明显的效果。进一步分析可以得到，在 $\alpha=0.1$ 时，指标FM的值比 $\alpha=0.7$ 时

FM值高，但是当 $\alpha=0.7$ 时，具有更好的AA和F指标。综合考虑，当 $\alpha=0.7$ 时，具有最优的聚类效果。IBOW与W2V组合的最优参数设置为 $\alpha=0.7$ ，$\beta=0.3$。

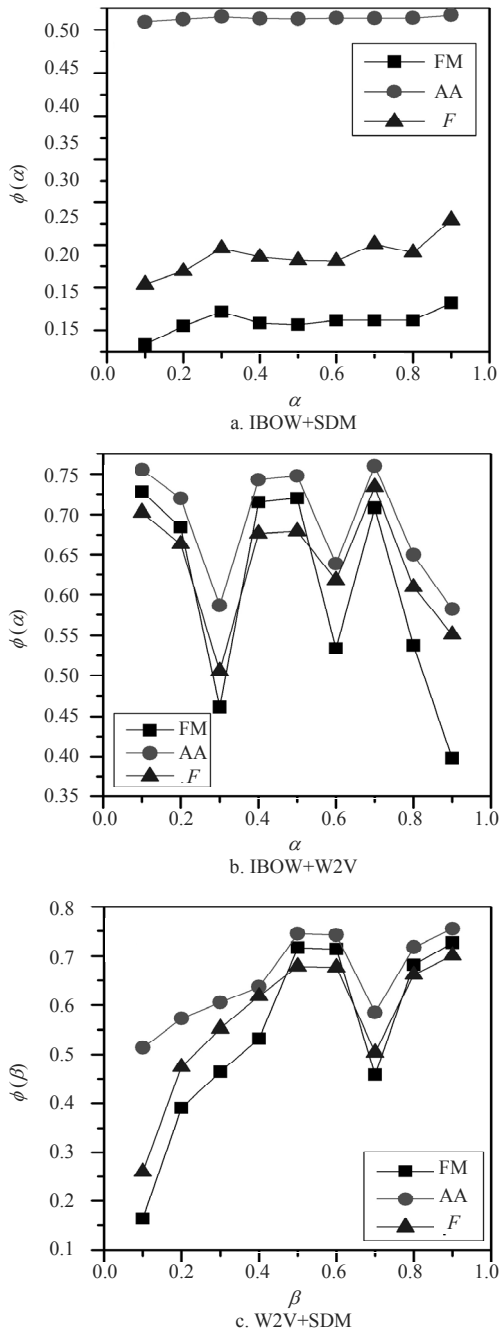

图1 两两组合模型的各项聚类指标对比

在第6组模型中，结合W2V和SDM对兴趣主题进行抽取，体现在式(7)中需要满足 $\alpha=0$ 。通过一组实验来观察参数 $\beta\in[0.1,1)$ 对各项指标值的影响，实验结果如图1c所示。从图1c可以看出，随着 β 的增加，各项指标FM、AA、F先增后减再增，当 $\beta=0.9$ 时，聚类指标FM、AA、F都达到最优值。因此W2V与SDM组合的最优参数设置为 $\alpha=0$ ， $\beta=0.9$ 。

在最后一组模型中，在IBOW的基础上，结合W2V和SDM对兴趣主题进行抽取，体现在式(7)中需要满足 $\alpha+\beta<1$ 。通过一组实验，分析聚类的各项指标FM、AA、F随着 α,β 的变化情况，其中 $\alpha,\beta\in[0.1,0.9]$ ，实验结果如图2所示。从图2a和2b可以看出，当 $\alpha\in[0.2,0.6],\beta=0.3$ 时，该模型的AA指标和F指标是白色区域并且达到最优。从图2c可以看出，当 $\alpha\in[0.2,0.3],\beta=0.3$ 时，该模型的FM指标是白色区域并且达到最优。综合考虑，IBOW+W2V+SDM模型的最优参数取为 $\alpha=0.2$ ， $\beta=0.3$ 。

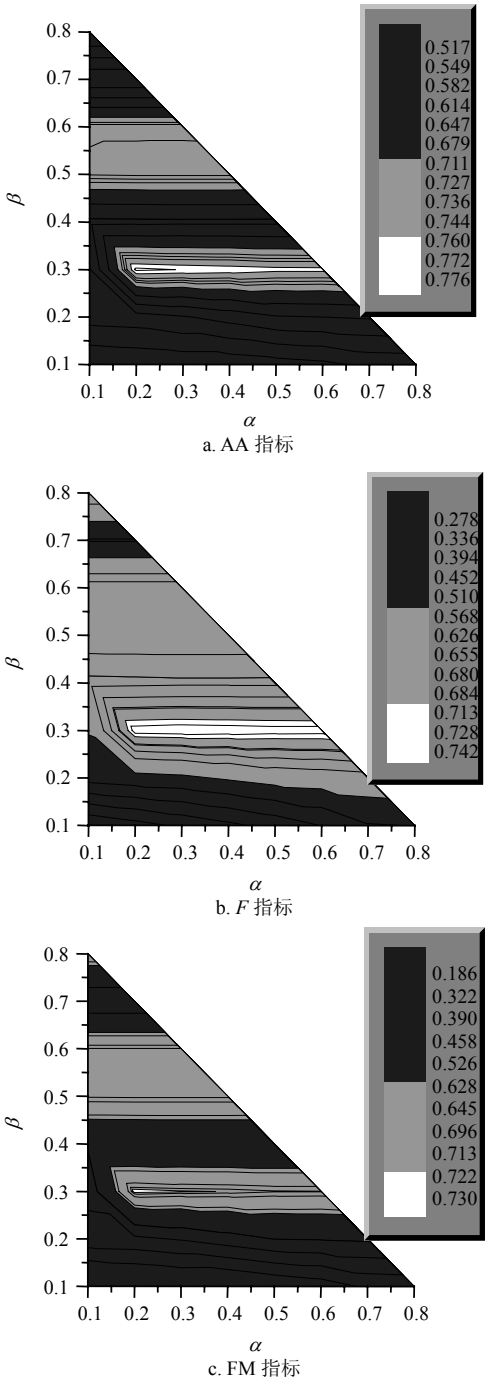

图2 IBOW+W2V+SDM模型的FM、AA、F指标

2) 不同阈值下各项聚类指标对比实验

保持IBOW+W2V+SDM模型的最优参数配置 $\alpha=0.2,\beta=0.3$，通过实验分析不同聚类融合条件阈值 $\theta\in[0,1)$ 下的各项指标情况，实验结果如图3所示。从图3可以看出，当 θ 从0.0逐渐增加到0.1时，各项聚类指标达到最大值，聚类效果达到最优，之后随着 θ 的增加各项聚类指标逐渐降低，这是因为随着 θ 的增加，聚类融合条件更严格，形成新话题的个数增加，而有效地类簇降低，导致整体聚类效果变差。

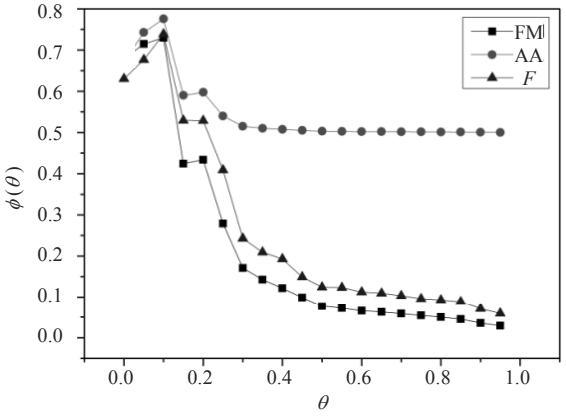

图3　不同阈值下各项聚类指标对比

3) 不同模型的对比

对7个不同模型做进一步比较分析，将每个模型的最好情况整理到一起，所得结果如图4所示。可以看出，IBOW+W2V+SDM模型的各项指标均具有最好的效果。其他组合模型中只要有W2V的加入，最后得到的结果都会更好，因此W2V在短文本的文本表示模型和相似度计算方面具有重要的作用。此外，IBOW+W2V+SDM模型在微博的兴趣主题抽取方面具有最佳的效果。

在Single-Pass聚类算法中，分为有质心更新和无质心更新两种情况，本文的主题抽取方法采用的是无质心更新的Single-Pass算法。并且与最新的方法FSC-LDA[15]进行了对比，具体的实验结果如表4所示。

从表4可以得到，IBOW能够在一定程度上提高聚类的各项指标。同时采用无质心更新能够得到更好的聚类效果。IBOW的FM、AA和*F*等聚类指标相比BOW分别提高了66.70%、2.17%和100.20%；IBOW+W2V+SDM(有质心更新)的FM、AA和*F*相比BOW分别提高了400.65%、26.28%和300.15%；IBOW+W2V+SDM(无质心更新)的FM、AA和*F*相比BOW分别提高了500.79%、53.75%和400.21%。本文基于Single-Pass算法的用户兴趣主题模型的FM、AA和*F*相比最新方法FSC-LDA分别提高了200.40%、46.50%、80.05%。

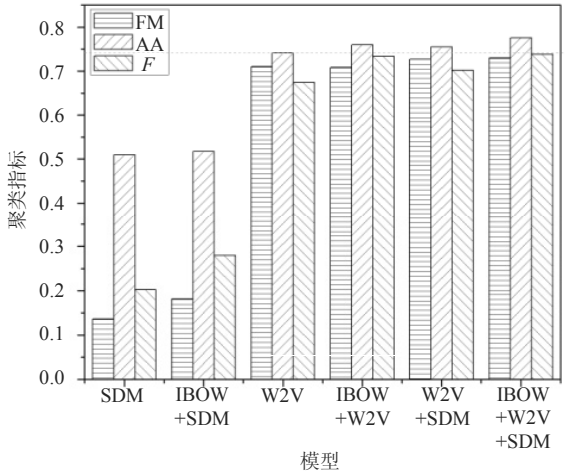

图4　不同模型下各项聚类指标对比

表4　模型的聚类效果比较

模型	FM	AA	*F*
BOW	0.108	0.506	0.142
IBOW	0.180	0.517	0.287
FSC-LDA	0.244	0.531	0.411
IBOW+W2V+SDM(有质心更新)	0.547	0.639	0.589
IBOW+W2V+SDM(无质心更新)	0.733	0.778	0.740

3　结 束 语

本文提出了一种微博用户兴趣主题抽取模型用于挖掘用户的兴趣。在该模型中提出了基于微博短文本特征的IBOW模型，基于语义关系的W2V模型，基于句子的顺序关系的SDM模型，综合考虑IBOW、W2V和SDM提取的特征，利用Single-Pass进行聚类，从而对用户兴趣进行准确的提取和更新。实验结果表明本文模型相比FSC-LDA方法具有更好的效果。

本文发现在IBOW+W2V模型和W2V+SDM模型中，各项聚类指标随着参数存在抖动现象，其原因有待进一步分析研究。下一步工作将考虑IBOW、W2V与SDM的组合与其他聚类算法进行融合的研究与应用。

参 考 文 献

[1] 周涛, 盛杨燕. 大数据时代[M]. 杭州: 浙江人民出版社, 2014.
ZHOU Tao, SHENG Yang-yan. Big data age[M]. Hangzhou: Zhejiang People's Publishing House, 2014.

[2] KWAK H, LEE C, PARK H, et al. What is Twitter, a social network or a news media[C]//Proceedings of the 19th International Conference on World Wide Web. [S.l.]: ACM, 2010.

[3] HA I, OH K J, JO G S. Personalized advertisement system using social relationship based user modeling[J]. Multimedia

Tools and Applications, 2015, 74(20): 8801-8819.
[4] DAO W, LE N, CHENG J, et al. Social media advertising value: the case of transitional economies in southeast Asia [J]. International Journal of Advertising, 2014, 33(2): 271-294.
[5] ANSTEAD N, O'LOUGHLIN B. Social media analysis and public opinion: the 2010 UK general election[J]. Journal of Computer-Mediated Communication, 2015, 20(2): 204-220.
[6] ERIKSON R S, TEDIN K L. American public opinion: Its origins, content and impact[M]. NewYork: Routledge, 2015.
[7] TURCOTTE J, YORK C, IRVING J, et al. News recommendations from social media opinion leaders: Effects on media trust and information seeking[J]. Journal of Computer-Mediated Communication, 2015, 20(5): 520-535.
[8] XIA H, YAN Z, BOWEN A. The mechanism and influencing factors of herding effect of college students' network public opinion[J]. Anthropologist, 2016, 23(1-2): 226-230.
[9] LI B, BAI B X, ZHANG C, et al. A method of network public opinion analysis based on quantum particle swarm algorithm optimization least square vector machine[J]. International Journal of Database Theory and Application, 2016, 9(8): 201-210.
[10] YU J, ZHU T. Combining long-term and short-term user interest for personalized hashtag recommendation[J]. Frontiers of Computer Science, 2015, 9(4): 608-622.
[11] FAN R, ZHAO J, XU K. Topic dynamics in Weibo: a comprehensive study[J]. Social Network Analysis and Mining, 2015, 5(1): 1-15.
[12] LI H, YAN J, HAN W, et al. Mining user interest in microblogs with a user-topic model[J]. Communications, China, 2014, 11(8): 131-144.
[13] LIU Q, NIU K, HE Z, et al. Microblog user interest modeling based on feature propagation[C]//2013 Sixth International Symposium on Computational Intelligence and Design (ISCID). [S.l.]: IEEE, 2013, 1: 383-386.
[14] BHATTACHARYA P, ZAFAR M B, GANGULY N, et al. Inferring user interests in the twitter social network[C]//Proceedings of the 8th ACM Conference on Recommender Systems. [S.l.]: ACM, 2014: 357-360.
[15] CHEN Y, LI W, GUO W, et al. Popular topic detection in Chinese micro-blog based on the modified lda model[C]//2015 12th Web Information System and Application Conference (WISA). [S.l.]: IEEE, 2015: 37-42.
[16] AN D, ZHENG X, RONG C, et al. Gaussian mixture model based interest prediction in social networks [C]//2015 IEEE 7th International Conference on Cloud Computing Technology and Science(CloudCom). [S.l.]: IEEE, 2015: 196-201.
[17] 格桑多吉, 乔少杰, 韩楠, 等. 基于Single-Pass的网络舆情热点发现算法[J].电子科技大学学报, 2015, 44(4): 599-604.
GESANG Duo-ji, QIAO Shao-jie, HAN Nan, et al. Network public opinion hot spot discovery algorithm based on Single-Pass[J]. Journal of University of Electronic Science and Technology, 2015, 44(4): 599-604.
[18] YAN T W, GARCIA-MOLINA H. Index structures for selective dissemination of information under the boolean model[J]. ACM Transactions on Database Systems (TODS), 1994, 19(2): 332-364.
[19] MARCU D, WONG W. A phrase-based, joint probability model for statistical machine translation[C]//Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing-Volume 10. [S.l.]: Association for Computational Linguistics, 2002: 133-139.
[20] BENGIO Y, DUCHARME R, VINCENT P, et al. A neural probabilistic language model[J]. Journal of Machine Learning Research, 2003, 3(2): 1137-1155.
[21] SALTON G, BUCKLEY C. Term-weighting approaches in automatic text retrieval[J]. Information Processing & Management, 1988, 24(5): 513-523.
[22] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space [EB/OL]. (2013-01-16). https://arxiv.org/abs/1301.3781.
[23] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[J]. Advances in Neural Information Processing Systems, 2013, 26: 3111-3119.
[24] 数据堂. 63 641个用户的新浪微博数据集[EB/OL]. [2015-03-24]. http://more.datatang.com/data/46758, 2014.
Data Church. 63 641 users of the Sina microblogging data set[EB/OL]. [2015-03-24]. http://more.datatang.com/data/46758,2014.
[25] 周昭涛. 文本聚类分析效果评价及文本表示研究[D]. 北京: 中国科学院研究生院(计算技术研究所), 2005.
ZHOU Zhao-tao. Text clustering analysis of the effect evaluation and text representation research[D]. Beijing: Graduate School of Chinese Academy of Sciences (Institute of Computing Technology), 2005.
[26] THEODORIDIS S, PIKRAKIS A, KOUTROUMBA K, et al. Introduction to pattern recognition: a Matlab approach [M]. [S.l.]: Academic Press, 2010.
[27] IWAYAMA M , TOKUNAGA T, Hierarchical Bayesian clustering for automatic text classification[C]//Proceedings of the 14th International Joint Conference on Artificial Intelligence-Volume 2. [S.l.]: Morgan Kaufmann Publishers Inc, 1995: 1322-1327.
[28] 姜宁, 宫秀军, 史忠植. 高维特征空间中文本聚类研究[J]. 计算机工程与应用, 2002, 38(10): 63-67.
JIANG Ning, GONG Xiu-jun, SHI Zhong-zhi. Study on text clustering in high dimensional feature space[J]. Computer Engineering and Applications, 2002, 38(10): 633-637.
[29] LARSEN B, AONE C. Fast and effective text mining using linear-time document clustering[C]//Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S.l.]: ACM, 1999: 16-22.

编 辑 蒋 晓